False Discovery Rate Control for Grouped Variable Selection in High-Dimensional Linear Models Using the T-Knock Filter

Jasin Machkour and Michael Muma
Signal Processing Group
Technische Universität Darmstadt
64283 Darmstadt, Germany
{j.machkour, muma}@spg.tu-darmstadt.de

Daniel P. Palomar
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
palomar@ust.hk

***Abstract*—High-dimensional variable selection is a challenging task, especially when groups of highly correlated variables are present in the data, such as in genomics research, direction-of-arrival estimation, and financial engineering. Recently, the T-Knock filter, a new framework for fast variable selection in high-dimensional settings has been developed. It provably controls the false discovery rate (FDR) at a given target level. However, its current version does not consider groups of highly correlated variables, which can lead to a loss in the true positive rate (TPR), i.e., the power. Hence, we propose the T-Knock+GVS filter that allows for grouped variable selection with FDR control in such settings. This is achieved by modifying the forward variable selection algorithm within the T-Knock filter and by adjusting the knockoff generation process such that the generated sets of knockoffs mimic the group correlation structure within the original set of variables. For a special case, we prove that the proposed T-Knock+GVS filter possesses the grouped variable selection property. Through a simulated high-dimensional genome-wide association study (GWAS), we show that the proposed method significantly increases the TPR, while controlling the FDR at the target level.**

***Index Terms*—T-Knock filter, grouped variable selection, false discovery rate (FDR) control, high-dimensional variable selection, genome-wide association studies (GWAS)**

I. Introduction

In many signal processing applications, the data is high-dimensional (i.e., more variables than data points/observations) and contains groups of highly correlated variables. For example, in genomics research, groups of nearby and, therefore, highly correlated single nucleotide polymorphisms (SNPs) occur throughout the genome due to a phenomenon called linkage disequilibrium [1]. In direction-of-arrival (DOA) estimation, groups of correlated signals are often present in the received signal at the sensor array [2]. In financial engineering, groups of highly correlated stocks from the same or related industries are usually present in the data [3]. All these applications can be modeled as variable selection tasks, in which it is necessary to select the right number of active variables (see, e.g., [4]–[6]).

Unfortunately, popular variable selection methods, such as the *Lasso* [7], *adaptive Lasso* [8], or even the *Elastic Net* [9] that is able to select groups of highly correlated variables, do not control the false discovery rate (FDR). The FDR is defined as the expected value of the false discovery proportion, i.e.,

$$\mathrm{FDR} := \mathbb{E}\big[\,\mathrm{FDP}\,\big] := \mathbb{E}\left[\frac{|\widehat{\mathcal{A}}\backslash\mathcal{A}|}{1 \vee |\widehat{\mathcal{A}}|}\right], \tag{1}$$

where $\widehat{\mathcal{A}} \subseteq \{1,\dots,p\}$ is the set of selected variables, $\mathcal{A} \subseteq \{1,\dots,p\}$ is the set of true active variables, $|\widehat{\mathcal{A}}|$ denotes the cardinality of the set $\widehat{\mathcal{A}}$, and $\vee$ is the maximum operator (i.e., $a \vee b = \max\{a,b\}$, $a,b \in \mathbb{R}$). A method has the FDR control property if it can determine $\widehat{\mathcal{A}}$ such that $\mathrm{FDR} \leq \alpha$, where $\alpha \in [0,1]$ is the user-defined target FDR level. The most popular FDR-controlling variable selection methods in low-dimensional settings are the Benjamini-Hochberg method [10] and the Benjamini-Yekutieli method [11]. In recent years, FDR-controlling methods for high-dimensional regression settings, such as the *model-X knockoff* methods [12] and the *T-Knock* filter [13], have been proposed.

In this paper, we build upon the *T-Knock* filter, whose computational complexity is linear in p, rendering it feasible for very high-dimensional data. The *T-Knock* filter also maximizes the number of selected variables, while maintaining FDR control. However, the *T-Knock* filter is not designed for grouped variable selection, which may lead to a decrease in the true positive rate (TPR), i.e., the power. The TPR is defined as the expected value of the true positive proportion (TPP), i.e.,

$$\mathrm{TPR} := \mathbb{E}\big[\,\mathrm{TPP}\,\big] := \mathbb{E}\left[\frac{|\mathcal{A} \cap \widehat{\mathcal{A}}|}{1 \vee |\mathcal{A}|}\right],$$

where $\widehat{\mathcal{A}}$ and $\mathcal{A}$ are defined as in (1).

In order to overcome this shortcoming, we propose the *T-Knock+Grouped Variable Selection* (*T-Knock+GVS*) filter,

The work of the first author has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant number 425884435. The work of the second author has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center and is supported by the 'Athene Young Investigator Programme' of Technische Universität Darmstadt, Darmstadt, Germany. The work of the third author has been funded by the Hong Kong GRF 16207820 research grant.

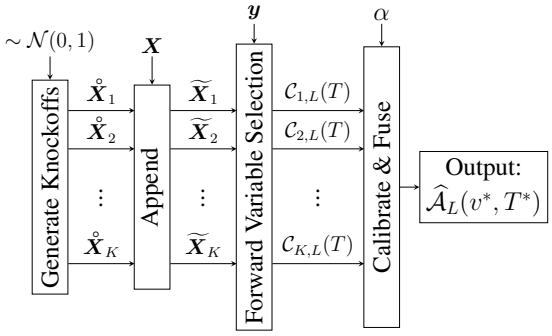


Figure 1: A simplified schematic overview of the *T-Knock* framework.

which integrates the Elastic Net as a forward variable selection method into the *T-Knock* framework (see Figure 1). For a special case, we prove that the proposed method possesses the grouped variable selection property. The *T-Knock+GVS* filter requires to design knockoff sets that mimic the group correlation structure of the original variables. Therefore, we propose a new knockoff generation process that generates suitable knockoff sets consisting of groups of correlated knockoffs.

The remainder of this paper is organized as follows: Section II briefly summarizes the existing *T-Knock* filter. In Section III, the proposed *T-Knock+GVS* filter is introduced and discussed. Section IV, compares the proposed method to benchmark methods via a simulated GWAS and Section V concludes the paper.

An implementation of the proposed T-Knock+GVS filter is available at https://github.com/jasinmachkour/tknock.

II. The T-Knock Filter

The Terminating-Knockoff (*T-Knock*) filter is a fast and FDR controlling variable selection framework for large-scale and high-dimensional as well as low-dimensional linear regression settings [13]. It considers the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_p] \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^p$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ with $\boldsymbol{0}$ being a vector of zeros and $\boldsymbol{I}$ being the identity matrix, are the response vector, the predictor matrix, the to be estimated coefficient vector, and the additive Gaussian noise with variance σ^2, respectively. Following the notation of the linear regression model, where p is the number of variables and n is the number of data points, all settings for which $p > n$ ($p \leq n$) are called high-dimensional (low-dimensional) settings. Moreover, it is assumed that the support of the coefficient vector $\boldsymbol{\beta}$ is sparse, i.e., only a few coefficients are non-zero, and that the predictors are standardized, i.e., $\sum_{i=1}^n x_{ij} = 0$, where x_{ij} is the ith element of $\boldsymbol{x}_j$, and $\|\boldsymbol{x}_j\|_2 = 1$ for all $j \in \{1, \dots, p\}$.

As illustrated in Figure 1, the *T-Knock* filter generates K knockoff matrices $\mathring{\boldsymbol{X}}_k \in \mathbb{R}^{n \times L}$, $k = 1, \dots, K$, each containing L knockoffs (i.e., fake predictors) that are sampled from the univariate standard normal distribution. These knockoff matrices are appended to the original predictor matrix $\boldsymbol{X}$, which yields the so-called enlarged predictor matrices $\widetilde{\boldsymbol{X}}_k = \left[\boldsymbol{X}\ \mathring{\boldsymbol{X}}_k\right]$, $k = 1, \dots, K$. The enlarged predictor matrices are used to conduct K independent random experiments. The random experiments are designed such that the knockoff variables compete with the given candidate variables in $\boldsymbol{X}$ to be included by a forward variable selection method, such as the *LARS* algorithm [14] or the closely related *Lasso* [7]. In each random experiment, the solution path is terminated early, as soon as a predefined number of $T \geq 1$ knockoffs is included by the forward variable selection method. This results in the K candidate sets $\mathcal{C}_{1,L}(T), \dots, \mathcal{C}_{K,L}(T)$ that contain all candidate variables that have been included before terminating the inclusion process after T knockoffs are included. The early stopping leads to a drastic reduction in computation time for sparse problems, where continuing the inclusion process leads to including more null variables. Finally, a calibration and fusion scheme that takes into account the user-defined target FDR level $\alpha \in [0, 1]$ and the relative occurrences of the candidate variables, denoted by $\Phi_{T,L}(j)$, $j \in \{1, \dots, p\}$, is applied to determine the optimal selected active set $\widehat{\mathcal{A}}_L(v^*, T^*)$, whose general definition is given by

$$\widehat{\mathcal{A}}_L(v, T) \coloneqq \{j : \Phi_{T,L}(j) > v\}. \tag{3}$$

We omit the details of how the optimal voting level $v^* \in [0.5, 1)$ and the optimal number of included knockoffs $T^* \geq 1$ are determined such that the FDR is provably controlled at the user-defined target FDR level while maximizing the number of selected variables. For these details, we refer the interested reader to the original *T-Knock* paper [13]. Moreover, the authors of the *T-Knock* filter propose an extended calibration algorithm that also determines the number of knockoffs L such that the FDR is more tightly controlled at low target levels.

III. Proposed Method: The T-Knock+GVS Filter

We propose the *T-Knock+GVS* filter, a grouped variable selection method that empirically controls the FDR. The proposed method has two major innovations that distinguish it from the original *T-Knock* filter:

A. It replaces the originally used variable selection method (*LARS* or *Lasso*) by the *Elastic Net*, which renders it suitable for performing grouped variable selection.

B. This replacement requires a major adjustment of the knockoff generation process. A new knockoff generation process that mimics the group correlation structure of $\boldsymbol{X}$ is proposed. This necessary adjustment fosters the generation of groups of highly correlated knockoffs that allow for a fair competition of original variables and knockoffs to be included along the forward variable selection process within each random experiment.

A. T-Knock+GVS Filter: Grouped Variable Selection

The naive *Elastic Net* combines the Lasso and Ridge regression. Its solution vector is given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1, \tag{4}$$

where $\lambda_1, \lambda_2 > 0$ are the weights for the sparsity inducing ℓ_1-norm penalty (Lasso) and the grouped selection fostering ℓ_2-norm penalty (Ridge regression), respectively. Since we are interested in performing grouped variable selection, we require a sufficiently large value of λ_2, such that the grouping effect is sufficiently strong. However, since the strength of the grouping effect is not very sensitive to the choice of λ_2, throughout this paper we will choose λ_2 by performing 10-fold cross validated Ridge regression and fix the obtained λ_2-value. With a fixed λ_2, the *Elastic Net* optimization problem can be reformulated as a *Lasso optimization problem* [9]. So, it can be solved by the *LARS* algorithm and it, therefore, can be integrated into the original *T-Knock* filter.

In the following, we will prove for the special case, where the variables within a group are perfectly correlated, that the desirable grouped variable selection property of the *Elastic Net* carries over to the proposed *T-Knock+GVS* filter. Considering this idealized case is common in theory, since it reveals whether a method is generally capable of performing grouped variable selection [9]. First, for each standardized variable $m \in \{1, \dots, p\}$, we define a group of perfectly correlated variables $\mathcal{G}_m$ that contains variable m. Then, we show that if any variable contained in $\mathcal{G}_m$ is selected (not selected) by the *T-Knock+GVS* filter, then the entire group is selected (not selected).

Theorem 1 (Grouped variable selection)**.** *Define* $\rho_{g,m} \coloneqq \boldsymbol{x}_g^\top \boldsymbol{x}_m$ *and*

$$\mathcal{G}_m \coloneqq \big\{g \in \{1, \dots, p\} : |\rho_{g,m}| = 1\big\}, m = 1, \dots, p. \quad (5)$$

The following two statements hold for all triples $(v, T, L) \in [0.5, 1) \times \{1, \dots, L\} \times \mathbb{N}_+$*:*

(i) Suppose that $j \in \mathcal{G}_m$ *and* $j \in \widehat{\mathcal{A}}_L(v, T)$*. Then, it holds that* $\mathcal{G}_m \subseteq \widehat{\mathcal{A}}_L(v, T)$*.*

(ii) Suppose that $j \in \mathcal{G}_m$ *and* $j \notin \widehat{\mathcal{A}}_L(v, T)$*. Then, it holds that* $\mathcal{G}_m \cap \widehat{\mathcal{A}}_L(v, T) = \varnothing$*, where* $\varnothing$ *denotes the empty set.*

Proof. The proof is deferred to the appendix. □

B. T-Knock+GVS Filter: Knockoff Generation Process

The goal of the proposed knockoff generation algorithm is to generate knockoff matrices $\mathring{\boldsymbol{X}}_k$, $k = 1, \dots, K$, that mimic the group correlation structure that is present within $\boldsymbol{X}$.

First, in order to cluster the variables into groups of highly correlated variables with low correlations between variables from different clusters, we apply single-linkage hierarchical clustering [15] to the predictors in $\boldsymbol{X}$, where the sample correlation is used as the similarity measure. Then, the obtained dendrogram is cut at the lowest level where the sample correlations of any two predictors from different clusters are not higher than the threshold value $\rho_{\text{thr}} = 1/3$. The value of ρ_{thr} is determined empirically, such that the resulting clusters capture the characteristic group correlation structure of SNPs. Such a clustering approach was proposed to be used as an SNP clustering method in, e.g., the supplementary material of [16]. As specified in the extended calibration algorithm in [13] that determines the value of L (i.e., the number of knockoffs), L is a multiple of the number of predictors p. Thus, L/p sub-knockoff matrices that mimic the group correlation structure of $\boldsymbol{X}$ are generated and appended together to obtain the final knockoff matrices $\mathring{\boldsymbol{X}}_k$, $k = 1, \dots, K$.

The annotated pseudocode of the proposed *T-Knock+GVS* knockoff generation process for the generation of the kth knockoff matrix $\mathring{\boldsymbol{X}}_k$ is given in Algorithm 1.

Algorithm 1 *T-Knock+GVS* knockoffs

1. **Input**: $\boldsymbol{X}$, ρ_{thr}, L.
2. **Apply** single-linkage hierarchical clustering [15] to the predictors in $\boldsymbol{X}$ and **cut** the resulting dendrogram at the lowest level where the sample correlations of any two predictors from different clusters are not higher than ρ_{thr}. Result: Z clusters with associated disjoint variable index sets $\mathcal{J}_1, \dots, \mathcal{J}_Z \subseteq \{1, \dots, p\}$, where $\bigcup_{z=1}^{Z} \mathcal{J}_z = \{1, \dots, p\}$.
3. **For** $w = 1, \dots, w_{\max}$, where $w_{\max} \coloneqq \dfrac{L}{p}$, **do**:

3.1. **For** $z = 1, \dots, Z$ **do**:

i. **Compute** the sub-cluster covariance matrix

$$\boldsymbol{\Sigma}_z = \frac{1}{n-1} \boldsymbol{X}_{\mathcal{J}_z}^\top \boldsymbol{X}_{\mathcal{J}_z},$$

where $\boldsymbol{X}_{\mathcal{J}_z}$ is the sub-matrix of $\boldsymbol{X}$ that contains the predictors corresponding to $\mathcal{J}_z$.

ii. **Compute** the sub-knockoff matrix

$$\mathring{\boldsymbol{X}}_{z,w} = \begin{bmatrix} \mathring{\boldsymbol{x}}'^\top_{z,w,1} \\ \vdots \\ \mathring{\boldsymbol{x}}'^\top_{z,w,n} \end{bmatrix}, \quad \mathring{\boldsymbol{x}}'_{z,w,i} \sim \mathcal{N}\big(\boldsymbol{0}, \boldsymbol{\Sigma}_z\big),$$

where $\mathring{\boldsymbol{x}}'^\top_{z,w,i}$ is the ith row of $\mathring{\boldsymbol{X}}_{z,w}$.

4. **Output**: kth knockoff matrix

$$\mathring{\boldsymbol{X}}_k = \Big[\mathring{\boldsymbol{X}}_{1,1} \cdots \mathring{\boldsymbol{X}}_{Z,1} \quad \cdots \quad \mathring{\boldsymbol{X}}_{1,w_{\max}} \cdots \mathring{\boldsymbol{X}}_{Z,w_{\max}}\Big].$$

IV. Simulated Genome-Wide Association Study

The goal of a genome-wide association study (GWAS) is to detect genetic variants, so-called single nucleotide polymorphisms (SNPs), across the genotype (i.e., genetic material) that are associated with a phenotype (i.e., observable and/or measurable characteristic of the disease of interest). In order to keep the number of false positives low and, therefore, foster reproducible discoveries in GWAS, it is important to control the FDR at a low target level. Since nearby SNPs across the genome usually form groups of highly correlated SNPs (see Figure 2), the proposed *T-Knock+GVS* filter is a suitable variable selection method for GWAS. In the following, we present and discuss the results of a simulated GWAS.

A. Performance Metrics and Benchmark Methods

The performance of the proposed *T-Knock+GVS* filter and the benchmark methods are compared in terms of the FDR

Table I The proposed *T-Knock+GVS* filter has the highest TPR while controlling the FDR at the target level of 20%. This shows that its grouped variable selection property leads to an enhanced performance. However, its sequential computation time is higher than that of the benchmark methods because of the increased row-dimension of the enlarged predictor matrices $\widetilde{\boldsymbol{X}}_k$, $k = 1, \ldots, K$, when solving the associated Elastic Net optimization problems via the LARS algorithm (see [9] for details).

Methods	**Average FDP (in %)**	**Average TPP (in %)**	**Average sequential computation time (hh:mm:ss)**	**Average relative sequential computation time**
Proposed: *T-Knock+GVS*	16.66	**58.77**	00:05:55	24.41
T-Knock	4.40	47.25	00:00:14	1
model-X+	3.84	14.03	00:00:40	2.75
model-X	7.21	44.22	00:00:40	2.75

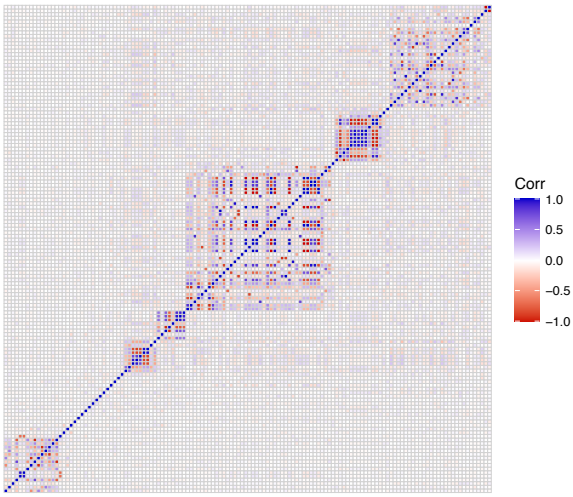
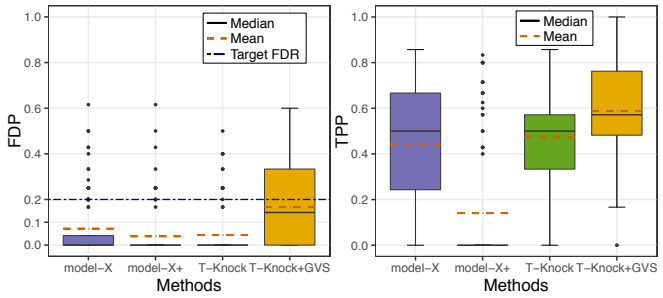


Figure 2: The heatmap visualizes the correlation matrix of 150 SNPs (containing three disease SNPs) that were generated using the software HAPGEN2 [17] and following the procedure that is described in Section IV-B.



(a) FDP-boxplots. (b) TPP-boxplots.

Figure 3: The proposed *T-Knock+GVS* filter has the highest TPR (i.e., average TPP), while its FDR (i.e., average FDP) stays below the target level of 20%. The benchmark methods do not fully take advantage of the target FDR level but stay significantly below it, which leads to a significantly lower TPR than the proposed method.

and TPR. The results show the averaged FDP and TPP over 100 Monte Carlo replications, which serve as estimates of the FDR and TPR, respectively.

We consider the following benchmark methods: original *T-Knock* filter [13], *model-X knockoff+* method [12], and *model-X knockoff* method [12].[1] The *model-X* methods generate and utilize knockoffs in a different way than the original *T-Knock* filter and the proposed *T-Knock+GVS* filter. For more details on the benchmark methods, we refer the interested reader to [12] and [13].

B. Setup and Results

Similar to the setup in [13], we simulate the genotypes of 700 study participants. Here, only the first 1,000 SNPs on Chromosome 15, of which 10 SNPs are associated with the disease of interest, are simulated using the software HAPGEN2 [17] and haplotypes from the International HapMap project (phase 3) [20]. The set of participants is divided into 500 cases and 200 controls. Since this is a case-control study, the phenotypes are binary, i.e., the case and control phenotypes are 1 and 0, respectively. The genotypes are represented by $\boldsymbol{X}$ in (2), where $n = 700$ is the number of study participants and $p = 1{,}000$ is the number of candidate SNPs. The phenotypes are represented by the response $\boldsymbol{y}$ in (2) with ones for cases and zeros for controls. The specific genomic parameters for the simulation of the data with HAPGEN2, such as the risk alleles, heterozygote risks, and homozygote risks, are chosen as in [13]. Also, the preprocessing of the data regarding the minor allele frequency, call rate, and Hardy-Weinberg disequilibrium is carried out as in [13]. However, in contrast to the preprocessing in [13], we do not carry out SNP pruning to reduce the dimension of the data but keep all SNPs.

Since the ultimate goal of a GWAS is to detect disease positions on the genome and not specific SNPs, it is reasonable to consider groups of highly correlated SNPs as active if they contain a disease SNP (see, e.g., [12], [16]). In this regard, a group of highly correlated SNPs is defined as a collection of SNPs of which no SNP has a correlation higher than $\rho_{\text{thr}} = 1/3$ with an SNP from another collection. The choice of ρ_{thr} is based on the same reasoning as in Algorithm 1. The results of the simulated GWAS are presented in Table I and Figure 3 and an interpretation thereof is given in the captions.

[1]Note that a group variable selection version of the fixed-X knockoff method [18] has been proposed in [19]. Unfortunately, however, it is designed for low-dimensional settings and can, therefore, not be considered as a benchmark method in our high-dimensional setting.

V. Conclusion

The *T-Knock+GVS* filter for FDR-controlled grouped variable selection in high-dimensional settings was proposed. Its FDR control property in the presence of groups of highly correlated variables was empirically verified. Moreover, it outperformed existing methods in terms of the TPR (i.e., power) on a high-dimensional simulated GWAS. Therefore, we consider the proposed *T-Knock+GVS* filter to be a suitable method for performing FDR-controlled and grouped variable selection in high-dimensional settings.

Acknowledgements

Extensive calculations on the Lichtenberg high-performance computer of the Technische Universität Darmstadt were conducted for this research.

Appendix

A. Proof of Theorem 1

Proof. First, note that, without loss of generality, $\mathcal{G}_m$ in (5) can be reduced to $\mathcal{G}_m = \big\{g \in \{1,\dots,p\} : \rho_{g,m} = 1\big\}, m = 1,\dots,p$, since x_g or x_m can be replaced by $-x_g$ or $-x_m$, respectively. The variable selection process in all K random experiments is not affected by such a replacement, because only the sign of the associated coefficient estimate is flipped.

Second, note that the relative occurrences within the definition of the selected active set in (3) are defined by

$$\Phi_{T,L}(j) \coloneqq \begin{cases} \dfrac{1}{K}\sum\limits_{k=1}^{K} \mathbb{1}_k(j,T,L), & T \geq 1 \\ 0, & T = 0 \end{cases},$$

where the indicator function in the first case is given by

$$\mathbb{1}_k(j,T,L) = \begin{cases} 1, & j \in \mathcal{C}_{k,L}(T) \\ 0, & \text{otherwise} \end{cases},$$

i.e., it is one if the jth variable is included in the candidate set of the kth random experiment and zero otherwise [13].

Third, note that Lemma 2 (a) in [9] states that for any strictly convex penalty function $f(\boldsymbol{\beta})$ in

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2 + \lambda f(\boldsymbol{\beta}),$$

it holds that if $\boldsymbol{x}_g = \boldsymbol{x}_m$, then $\hat{\beta}_g = \hat{\beta}_m$, $g,m \in \{1,\dots,p\}$, for all $\lambda > 0$. Since the elastic net penalty in (4) is a strictly convex function of $\boldsymbol{\beta}$, we conclude that $\rho_{g,m} = 1$ implies $\hat{\beta}_g = \hat{\beta}_m$, $g,m \in \{1,\dots,p\}$, for all $\lambda > 0$. From $\hat{\beta}_{g,k} = \hat{\beta}_{m,k}$ for all $\lambda > 0$, where $\hat{\beta}_{g,k}$ and $\hat{\beta}_{m,k}$ are the coefficient estimates of variables $\boldsymbol{x}_g$ and $\boldsymbol{x}_m$ corresponding to the kth random experiment, it follows that

$$\mathbb{1}_k(g,T,L) = \mathbb{1}_k(m,T,L)$$

for all $k \in \{1,\dots,K\}$ and all tuples $(T,L) \in \{1,\dots,L\} \times \mathbb{N}_+$. Consequently, $\rho_{j,m} = 1$ implies $\Phi_{T,L}(j) = \Phi_{T,L}(m)$ for all $j \in \mathcal{G}_m$. Thus, for all triples $(v,T,L) \in [0.5,1) \times \{1,\dots,L\} \times \mathbb{N}_+$ and for all $j,m \in \{1,\dots,p\}$, the following two statements hold:

(a1) If $\rho_{j,m} = 1$ and $\Phi_{T,L}(j) > v$, then $\Phi_{T,L}(m) > v$.
(b1) If $\rho_{j,m} = 1$ and $\Phi_{T,L}(j) \leq v$, then $\Phi_{T,L}(m) \leq v$.

Using the definition of $\widehat{\mathcal{A}}_L(v,T)$ in (3), Statements (a1) and (b1) can be translated into the following equivalent statements that hold for all triples $(v,T,L) \in [0.5,1) \times \{1,\dots,L\} \times \mathbb{N}_+$ and for all $j,m \in \{1,\dots,p\}$:

(a2) If $j \in \mathcal{G}_m$ and $j \in \widehat{\mathcal{A}}_L(v,T)$, then $G_m \subseteq \widehat{\mathcal{A}}_L(v,T)$.
(b2) If $j \in \mathcal{G}_m$ and $j \notin \widehat{\mathcal{A}}_L(v,T)$, then $G_m \cap \widehat{\mathcal{A}}_L(v,T) = \varnothing$.

Statements (a2) and (b2) are equivalent to Statements (i) and (ii) in the theorem. □

References

[1] D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward *et al.*, "Linkage disequilibrium in the human genome," *Nature*, vol. 411, no. 6834, pp. 199–204, 2001.

[2] T.-J. Shan, M. Wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation of coherent signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 4, pp. 806–811, 1985.

[3] R. N. Mantegna and H. E. Stanley, *An introduction to econophysics: Correlations and complexity in finance*. Cambridge Univ. Press, 1999.

[4] A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis *et al.*, "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1005–D1012, 2019.

[5] Z. Tan, Y. C. Eldar, and A. Nehorai, "Direction of arrival estimation using co-prime arrays: A super resolution viewpoint," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5565–5576, 2014.

[6] K. Benidis, Y. Feng, and D. P. Palomar, "Sparse portfolios for high-dimensional financial index tracking," *IEEE Trans. Signal Process.*, vol. 66, no. 1, pp. 155–170, 2017.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[8] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[9] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[10] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 57, no. 1, pp. 289–300, 1995.

[11] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann. Statist.*, vol. 29, no. 4, pp. 1165–1188, 2001.

[12] E. J. Candès, Y. Fan, L. Janson, and J. Lv, "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 80, no. 3, pp. 551–577, 2018.

[13] J. Machkour, M. Muma, and D. P. Palomar, "The Terminating-Knockoff filter: Fast high-dimensional variable selection with false discovery rate control," *arXiv preprint arXiv:2110.06048*, 2021. [Online]. Available: https://arxiv.org/abs/2110.06048

[14] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[15] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 1, pp. 86–97, 2012.

[16] M. Sesia, C. Sabatti, and E. J. Candès, "Gene hunting with hidden markov model knockoffs," *Biometrika*, vol. 106, no. 1, pp. 1–18, 2019.

[17] Z. Su, J. Marchini, and P. Donnelly, "HAPGEN2: simulation of multiple disease SNPs," *Bioinformatics*, vol. 27, no. 16, pp. 2304–2305, 2011.

[18] R. F. Barber and E. J. Candès, "Controlling the false discovery rate via knockoffs," *Ann. Statist.*, vol. 43, no. 5, pp. 2055–2085, 2015.

[19] R. Dai and R. Barber, "The knockoff filter for FDR control in group-sparse and multitask regression," in *33rd Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1851–1859.

[20] The International HapMap 3 Consortium, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, no. 7311, pp. 52–58, 2010.